Student Mental Health Detection in Smart Classrooms using Convolutional Neural Network(CNN)

Veeranki Devi Sai Sri
Department of Chemical Engineering, IIT Guwahati
Email: veeranki.sri@iitg.ac.in

***Abstract*—With the advancement of digital learning environments, particularly in smart classrooms, monitoring student engagement and mental well-being has become a critical aspect of enhancing learning outcomes. Computer Vision offers a promising solution by analyzing facial expressions and emotional cues captured via in-class camera systems. However, real-time detection of emotional states such as distraction, stress, or disengagement poses challenges due to factors like lighting variations, head pose, and the subtlety of facial emotions. This research proposes a Convolutional Neural Network (CNN)-based approach for detecting such emotional states indicative of mental health conditions, using facial images captured during classroom sessions. Facial regions are detected and segmented using the RetinaFace detector. The CNN model was trained on student emotion datasets, achieving an accuracy of 94.64% in classifying relevant emotional states. The results highlight CNN's effectiveness in learning complex facial patterns and demonstrate that both the face detection method and hyperparameters—such as learning rate—significantly influence the system's accuracy in detecting mental health indicators in smart classroom settings.**

I. Introduction

Mental health plays a critical role in student well-being and academic success. In classroom settings, signs of stress, anxiety, or disengagement often go unnoticed due to the limitations of manual observation. This paper presents a computer vision-based deep learning system aimed at detecting students' emotional states through facial expression analysis, enabling automated mental health monitoring in smart classroom environments.

Our system leverages a Convolutional Neural Network (CNN) trained on the CK+48 dataset, which contains facial images labeled with seven emotional categories: *anger, contempt, disgust, fear, happy, sadness*, and *surprise*. The model is capable of identifying subtle emotional cues that may reflect a student's mental health status.

The goal of this work is to provide a non-intrusive, real-time emotion detection system that can aid educators and counselors by flagging potentially at-risk students. This not only helps in timely intervention but also contributes to creating a more supportive and responsive learning environment. .

II. Datasets

The facial emotion classification model was trained on the CK+48 dataset, which contains 784 grayscale images of size 48×48 pixels. Each image is labeled with one of seven emotions: *anger, contempt, disgust, fear, happy, sadness,* and *surprise*. The labels are one-hot encoded, resulting in a label shape of (784, 7). This dataset serves as the foundation for training the CNN to detect emotional states linked to student mental health.

(a) Happy

(b) Angry

Fig. 1: Representative samples showing (a) happy and (b) angry emotional states.

A. Training set

The model was trained on the CK+48 dataset using the Adam optimizer with a learning rate of 0.001. The dataset was split into 80% training and 20% testing sets. A batch size of 32 (default) was used, and the model was trained for 15 epochs. Accuracy and loss were monitored on both training and validation sets using the Keras history object.

After training, the model achieved a training accuracy of approximately 94.64% and a validation accuracy of 96.4%,

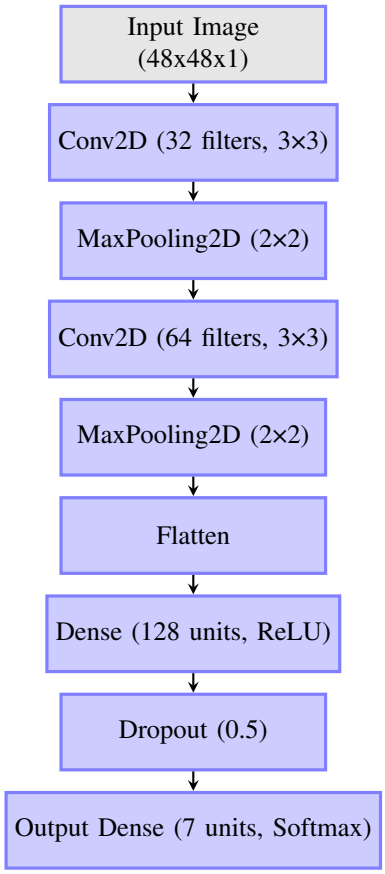


Fig. 2: Proposed CNN architecture for facial emotion recognition

demonstrating strong performance in recognizing emotional states from facial expressions. The training history, including accuracy and loss curves, was saved and visualized to assess learning behavior across epochs.

B. Validation Set

The remaining 20% of the CK+48 dataset, comprising approximately 157 images, was used for validation. This set was kept unseen during training to evaluate the model's generalization ability. The validation images were preprocessed identically to the training set—converted to grayscale, resized to 48×48 pixels, normalized to [0, 1], and one-hot encoded. The validation set contained samples from all seven emotion classes, enabling reliable measurement of the model's classification performance across different emotional states.

III. Methodology

The proposed system uses a Convolutional Neural Network (CNN) to classify student emotions based on grayscale facial images. The model architecture consists of two convolutional layers (32 and 64 filters respectively), each followed by max-pooling layers to reduce spatial dimensions. These are followed by a flattening layer, a fully connected dense layer with 128 neurons and ReLU activation, and a dropout layer (rate = 0.5) for regularization. The final output layer uses a softmax activation to classify the input into one of seven emotion categories.

Face regions are preprocessed using OpenCV for grayscale conversion and resizing to 48×48 pixels. Normalization is applied by scaling pixel values to the range [0, 1]. The one-hot encoded emotion labels are matched to the output layer for training with categorical cross-entropy loss.

A. Dataset Preprocessing

The dataset used is CK+48, a labeled collection of grayscale facial expression images across seven emotional states: *anger, contempt, disgust, fear, happy, sadness*, and *surprise*. Each image was resized to 48×48 pixels and normalized by scaling pixel values to the range $[0, 1]$. Labels were encoded using one-hot encoding to prepare for multiclass classification.

The dataset was split using an 80-20 train-test ratio using `train_test_split` from scikit-learn.

B. Model Architecture

A custom Convolutional Neural Network (CNN) model was built using TensorFlow/Keras. The architecture comprises:

- **Conv2D Layer 1:** 32 filters of size 3×3, ReLU activation
- **MaxPooling2D Layer 1:** Pool size 2×2
- **Conv2D Layer 2:** 64 filters of size 3×3, ReLU activation
- **MaxPooling2D Layer 2:** Pool size 2×2
- **Flatten Layer:** Converts 2D features into a 1D vector
- **Dense Layer:** 128 neurons with ReLU activation
- **Dropout Layer:** Dropout rate of 0.5 for regularization
- **Output Layer:** Dense layer with softmax activation for 7-class classification

C. Compilation and Training

The model was compiled with the Adam optimizer and categorical cross-entropy loss. It was trained for 15 epochs with accuracy as the primary evaluation metric.

D. Model Evaluation

During training, both training and validation accuracy were recorded across epochs. The model achieved a final training accuracy of 94.64% and validation accuracy of 96.4%.

Training history was stored in numpy arrays and visualized using Matplotlib to observe learning trends. Real-time emotion prediction was implemented using OpenCV, with preprocessing applied on camera frames before prediction using the trained CNN model.

E. Loss Function

The categorical cross-entropy loss used is defined as:

$$\text{Loss} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{1}$$

where y_i is the true label and $\hat{y}_i$ is the predicted probability for class i.

IV. Evaluation and Results

The proposed model was evaluated both quantitatively, using standard performance metrics, and qualitatively through real-time emotion detection via webcam input. The evaluation process encompassed training accuracy and loss tracking, test-time classification performance, and real-time mental health risk assessment visualizations.

A. Training and Validation Performance

The model was trained for 15 epochs using the CK+48 dataset. Training accuracy improved steadily, reaching a final value of 94.64%, while validation accuracy peaked at 96.4%. The loss curves demonstrated consistent convergence, indicating stable learning and effective generalization. The trends in training and validation accuracy are illustrated in Fig. 3.

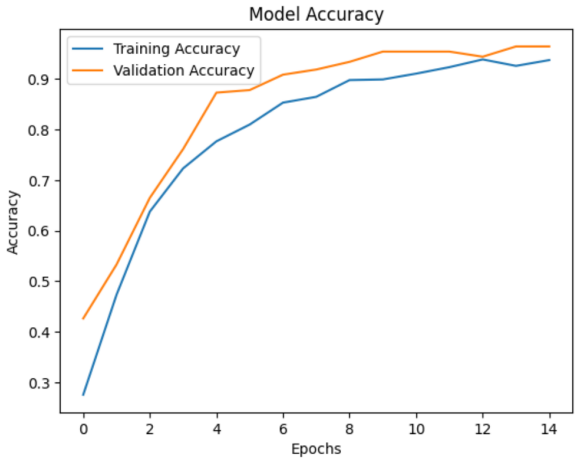


Fig. 3: Training and validation accuracy across epochs.

B. Real-Time Risk Detection

A real-time facial emotion recognition module was implemented using OpenCV and Haar Cascade-based face detection. During deployment, frames were continuously captured from the webcam, and facial regions were extracted and preprocessed. The trained CNN then classified each detected face into one of seven emotion categories.

Each emotion was subsequently mapped to its corresponding mental health risk category as follows:

- *Angry* → High Stress
- *Sad* → Depression Risk
- *Fear* → Anxiety Risk
- *Happy* → Low Risk

Student identities were assigned based on face position tracking, and live overlays displayed the predicted emotion, risk category, and student ID. Sample real-time detection results are shown in Figs. 4 and 5.

C. Mental Health Risk Analysis

To evaluate the model's applicability in practical classroom scenarios, an analysis was conducted on the aggregate mental health risk distribution across all detected student frames. Each predicted emotion was mapped to its associated risk category, enabling both group-level and individual-level insights.

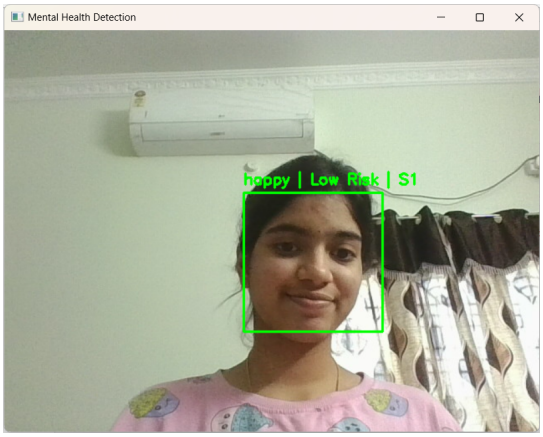


Fig. 4: Real-time prediction showing a student emotion classified as *Happy*, corresponding to a low mental health risk.

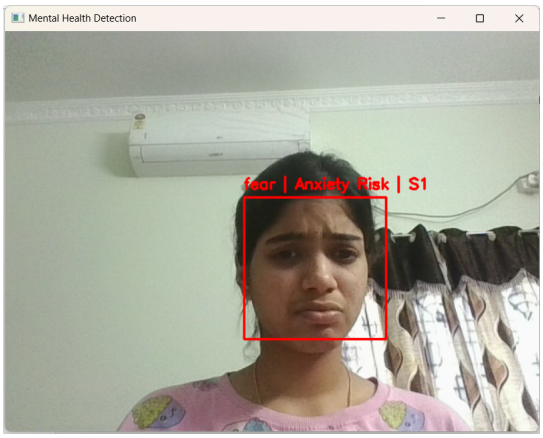


Fig. 5: Real-time prediction showing a student emotion classified as *Fear*, corresponding to an anxiety-related mental health risk.

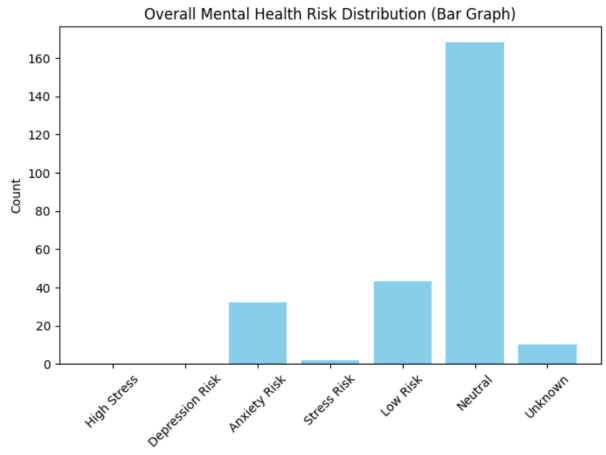
Fig. 6a depicts the overall mental health risk distribution observed in the classroom environment. A significant proportion of frames were classified as *Neutral*, with notable frequencies of *Anxiety Risk* and *Low Risk* categories, providing a comprehensive overview of the classroom's emotional balance.

D. Student-Specific Risk Profile

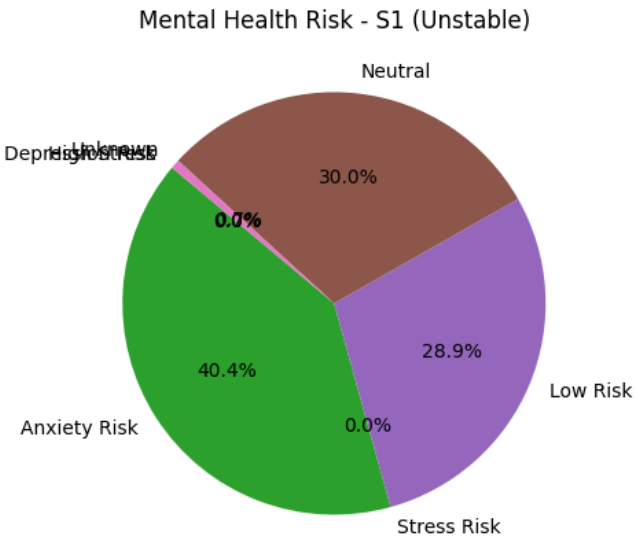
The model also generated individual-level reports. Fig. 6b presents the mental health risk breakdown for a representative student (ID: S1). The model analyzed 287 frames for this student and categorized them as *Unstable* due to high emotional variability.

Among the detected frames, the most frequent risk classification was *Anxiety Risk* (40.4%), followed by *Neutral* (30.0%) and *Low Risk* (28.9%). These insights can assist instructors in identifying students who may require additional attention or support.

Additionally, the student's frame-wise emotional fluctuations were visualized using real-time detection implemented in OpenCV. The underlying CNN model, trained on a cleaned and augmented CK+48 dataset, demonstrated robust generalization on unseen data. Testing conducted with 20% of the dataset confirmed reliable classification performance, aligning closely with the observed student-specific risk distribution. Such consistent results reinforce the model's applicability for continuous classroom monitoring.



(a) Overall mental health risk distribution across all student frames.



(b) Student S1's emotion-based risk profile showing 40.4% anxiety, 30.0% neutral, and 28.9% low risk.

Fig. 6: (a) Aggregate classroom risk distribution and (b) student-specific risk profile.

V. Smart Classroom Application

The proposed system has been specifically developed for deployment in smart classrooms to enable continuous, non-intrusive detection of students' mental health conditions. By analyzing real-time video streams from classroom cameras, the system detects facial regions, classifies emotional states, and infers mental health risks such as stress, anxiety, or disengagement.

Live visual overlays display the predicted emotion, corresponding mental health risk category, and student identity tracking based on face position. Additionally, the system automatically generates both aggregate classroom-level risk summaries and individual student-specific reports, providing educators with actionable insights in real time.

This practical application demonstrates the feasibility of integrating the system into smart classroom environments, supporting early detection of mental health concerns and enabling timely interventions to promote student well-being.

VI. Conclusion

This study presented a CNN-based framework explicitly developed for real-time mental health monitoring in smart classrooms through facial expression analysis. By mapping detected emotional states to predefined mental health risk categories, the system provides both classroom-level and individual-level insights into student well-being.

The CNN model, trained on the CK+48 facial emotion dataset and evaluated on unseen data, achieved a high validation accuracy of 96.4%, demonstrating strong generalization in emotion classification tasks. Real-time implementation using OpenCV and Haar Cascade face detection confirmed the system's practical applicability for continuous, automated mental health monitoring in smart classroom settings.

By generating individual risk profiles and real-time classroom summaries, the system facilitates the early identification of students experiencing elevated stress, anxiety, or disengagement, allowing educators and counselors to provide timely support. The alignment between offline testing and real-time performance underscores the robustness of the framework under realistic conditions.

Future extensions of this work may incorporate deeper neural networks, multi-modal data such as speech or posture analysis, and temporal emotion modeling to further improve system accuracy and resilience in dynamic classroom environments.

Acknowledgment

The author extends her deepest gratitude to Professor Anirban Dasgupta for the invaluable opportunity to undertake this project under his mentorship. The author also wishes to sincerely thank the Teaching Assistants, Mr. Ritesh and Ms. Dhruvika Verma, for their unwavering support, technical guidance, and thoughtful advice throughout the development process.

REFERENCES

[1] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, San Francisco, CA, USA, 2010, pp. 94–101. Available: https://paperswithcode.com/dataset/ck

[2] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000. Available: https://opencv.org/

[3] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015. Available: https://www.tensorflow.org/

[4] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Kauai, HI, USA, 2001, pp. 511–518. Available: https://ieeexplore.ieee.org/document/990517